The exact amount of t-ness that the normal model can tolerate

Nils Lid Hjort

University of Oslo and Norwegian Computing Centre

-- February 1991 --

ABSTRACT. Suppose that the normal model is used for data $Y_1, \ldots, Y_n$, but that the true distribution is a t-distribution with location and scale parameters ξ and σ and m degrees of freedom. The normal model corresponds to $m = \infty$. Using a local asymptotic framework where m is allowed to increase with n it is shown that if $m \geq 1.458\sqrt{n}$, then estimation using the incorrect normal model is still more precise than using the correct three-parameter model. This result is valid for all smooth parameter estimands, and is also true in regression models with t-distributed residuals. We also propose and analyse compromise estimators that interpolate smoothly between the normal and the nonnormal models. Proving our results requires somewhat nonstandard 'corner asymptotics' since behaviour of estimators must be studied when the crucial parameter $\gamma = 1/m$ is close to zero, which is not an inner point of the parameter space, and $\widehat{\gamma} = 0$ ($\widehat{m} = \infty$) with positive probability.

KEY WORDS: *choice of model, corner asymptotics, deliberate bias, guarding against heavier tails, ignorance is strength, misspecified model, negative degrees of freedom, parametric inference, tolerance radius*

1. Introduction and summary. The most popular model for independent identically distributed (i.i.d.) data $Y_1, \ldots, Y_n$ is to postulate normality, i.e. assuming $f(y) = \phi((y-\xi)/\sigma)/\sigma$ for suitable parameters ξ and σ. In many situations the normal density is too light-tailed to constitute a serious description, however. A remedy then is to use

$$f(y, \xi, \sigma, m) = g_m\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma},$$

where $g_m(t)$ is the t-density with m degrees of freedom. The narrower normal model corresponds to $m = \infty$, and it is naturally felt that for large m the discrepancy between normality and t-ness shouldn't matter. One would also expect inference methods based on the formally incorrect normality assumption to work better than inference methods based on the wider three-parameter model, for large values of m, since the estimation variability increases with the inclusion of m.

This paper studies the problem of choosing between 'narrow model' and 'wide model' estimators. Using the narrow method means introducing a certain bias due to incorrect modelling, whereas using the wide method means allowing additional parameter estimation noise. Choosing between the estimators therefore amounts to a statistical balancing act with perhaps deliberate bias against sampling variability.

Suppose for example that the parameter to be estimated is the upper quartile $q = F^{-1}(.75)$. Concentrating on maximum likelihood estimators the two methods amount to

$$\widehat{q}_{\rm narr} = \widehat{\xi}_{\rm narr} + .675\,\widehat{\sigma}_{\rm narr} \quad \text{and} \quad \widehat{q}_{\rm wide} = \widehat{\xi}_{\rm wide} + G^{-1}(.75, \widehat{m}_{\rm wide})\,\widehat{\sigma}_{\rm wide}.$$

Here $G(.,m) = G_m(.)$ is the t-distribution function with m degrees of freedom, with inverse $G^{-1}(.,m) = G_m^{-1}(.)$, whereas $\widehat{\xi}_{\rm narr}$ and $\widehat{\sigma}_{\rm narr}$ are the ML estimators under the narrow two-parameter model, and $\widehat{\xi}_{\rm wide}$, $\widehat{\sigma}_{\rm wide}$, $\widehat{m}_{\rm wide}$ are the ML estimators under the wide three-parameter model. [The narrow model estimators are of course the ordinary empirical mean and empirical standard deviation statistics, whereas ML estimates under the wide model must be computed by numerical maximisation techniques.] How large must m be in order for $\widehat{q}_{\rm narr}$ to be more precise than $\widehat{q}_{\rm wide}$? Suppose for a second example that the parameter to be estimated is sd, the standard deviation for Y_i's. We should compare

$$\widehat{\rm sd}_{\rm narr} = \Big\{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\Big\}^{1/2} \quad \text{and} \quad \widehat{\rm sd}_{\rm wide} = \sqrt{\frac{\widehat{m}_{\rm wide}}{\widehat{m}_{\rm wide}-2}}\,\widehat{\sigma}_{\rm wide}.$$

For what range of m values is the narrow estimator more precise than the much more laborious wide estimator, and for which m values will it be advantageous to use the latter? And a third example is that of estimating the probability $p(y) = \Pr\{Y_i \le y\}$, in which case the two estimators to be compared are

$$\widehat{p}_{\rm narr}(y) = \Phi\big((y - \widehat{\xi}_{\rm narr})/\widehat{\sigma}_{\rm narr}\big) \quad \text{and} \quad \widehat{p}_{\rm wide}(y) = G\big((y - \widehat{\xi}_{\rm wide})/\widehat{\sigma}_{\rm wide}, \widehat{m}_{\rm wide}\big).$$

These problems have a surprisingly sharp and simple solution. A natural large-sample framework is established in Section 2, where results are reached for the large-sample behaviour of ML estimators for (ξ,σ) in the narrow model and of ML estimators for (ξ,σ,γ) in the wide model, where $\gamma = 1/m$. These are used in Section 3 to solve the problem. It turns out that if only $m \ge 1.458\sqrt{n}$, then the narrow method is more precise than the wide method, in terms of mean squared errors, and this (large-sample) answer is valid for *all* parameter estimands. If t-ness is present with $m < 1.458\sqrt{n}$ then the wide method is more precise. Thus $1.458\sqrt{n}$ is effectively the 'tolerance distance' from the normal model w.r.t. t-ness.

A fuller story is told in Section 4, where a general class of compromise estimators is considered. These interpolate between the narrow normal and the wider t-model. We single out a few of these that are designed to work well both under normal and non-normal conditions. The regression case, where residuals can have a t distribution, is treated in Section 5. Results from the previous sections can be extended, and the tolerance distance becomes again precisely $1.458\sqrt{n}$. Our compromise estimators in this case can be viewed as a basis for performing robust regression analysis, guarding against heavier-than-normal tails. Finally some complementing remarks are offered in Section 6, including a construction of a quasi-t-distribution that allows negative degrees of freedom.

The problems about balancing modelling bias and estimation variability for incorrectly specified parametric models are obviously of a general nature, and can be studied for other important models as well as on a general basis. Such a study is indeed reported on in Hjort (1991), which contains further background, a general theory, and explicit results for a generous list of commonly used statistical models. The present t-ness departure case is however non-regular and cumbersome, and cannot be handled as a special case of the general regular theory. What makes this problem special is that the model must be studied when the crucial parameter $\gamma = 1/m$ is close to zero, which is not an inner point in the parameter space. In particular $\widehat{\gamma}_{\rm wide} = 0$ ($\widehat{m}_{\rm wide} = \infty$) with positive probability, the familiar ML asymptotics break down, and special methods are called for.

2. Large sample framework for the problem. The wide model has parameters ξ, σ, m. Let us reparameterise to $\gamma = 1/m$, so that the density becomes

$$f(y,\xi,\sigma,\gamma) = \frac{c(\gamma)}{\sigma}\Big\{1+\gamma\Big(\frac{y-\xi}{\sigma}\Big)^2\Big\}^{-\{1/2+1/(2\gamma)\}}, \quad c(\gamma) = \frac{\sqrt{\gamma}}{\sqrt{\pi}}\frac{\Gamma(\frac{1}{2}+\frac{1}{2\gamma})}{\Gamma(\frac{1}{2\gamma})}. \tag{2.1}$$

We are interested in this model for γ in the vicinity of zero. Using careful Taylor expansions and approximations to the $\log\Gamma(.)$ function one can show that

$$\log f(y,\xi,\sigma,\gamma) = \log f(y,\xi,\sigma,0) + \gamma(\tfrac{1}{4}z^4 - \tfrac{1}{2}z^2 - \tfrac{1}{4}) + \gamma^2(\tfrac{1}{4}z^4 - \tfrac{1}{6}z^6) + O(\gamma^3), \tag{2.2}$$

in which $z = (y-\xi)/\sigma$. Having $\gamma = 0$ corresponds to $m = \infty$ and gives back the ordinary normal model.

Let $\mu = \mu(f) = \mu(\xi,\sigma,\gamma)$ be some parameter estimand of interest. We assume that μ is smooth with continuous derivatives throughout the inner parameter space $(\xi,\sigma,\gamma) \in (-\infty,\infty)\times(0,\infty)\times(0,\infty)$ and that the right derivative exists at $\gamma = 0$, $\lim_{\gamma\to 0+}\{\mu(\xi,\sigma,\gamma) - \mu(\xi,\sigma,0)\}/\gamma$. We concentrate on ML procedures, and wish to study the performance of the two estimators

$$\widehat{\mu}_{\rm narr} = \mu(\widehat{\xi}_{\rm narr},\widehat{\sigma}_{\rm narr},0) \quad \text{and} \quad \widehat{\mu}_{\rm wide} = \mu(\widehat{\xi},\widehat{\sigma},\widehat{\gamma}), \tag{2.3}$$

where for simplicity of notation the subscript 'wide' is dropped for the ML estimators in the three-parameter model.

These could be compared in an asymptotic framework in which Y_i's come from some fixed $f(y,\xi_0,\sigma_0,\gamma)$, and $\gamma > 0$. In this case $\sqrt{n}(\widehat{\mu}_{\rm wide} - \mu)$ has a limit distribution. The situation is different for the narrow model procedure. Here $\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu)$ can be represented as a sum of two terms. The first is $\sqrt{n}\{\mu(\widehat{\xi}_{\rm narr},\widehat{\sigma}_{\rm narr},0) - \mu(\xi_0,\sigma_0,0)\}$, which has a limit distribution, with generally smaller variability than that of the wide model procedure; and the second is $-\sqrt{n}\{\mu(\xi_0,\sigma_0,\gamma) - \mu(\xi_0,\sigma_0,0)\}$, which tends to plus or minus infinity, reflecting a bias that for very large n will dominate completely. This merely goes to show that with very large sample sizes one is penalised for any bias and one should use the wide model. This result is somewhat irrelevant, however, and suggests that a large sample framework which uses a local neighbourhood of $\gamma = 0$ that shrinks when the sample size grows is much more adequate. Consider therefore model P_n, the n'th model, under which

$$Y_1,\ldots,Y_n \text{ are i.i.d. from } f_n(y) = f(y,\xi_0,\sigma_0,\delta/\sqrt{n}). \tag{2.4}$$

Here (ξ_0,σ_0) is a fixed but arbitrary parameter point. The true parameter to be estimated is $\mu_{\rm true} = \mu(\xi_0,\sigma_0,\delta/\sqrt{n})$. To assess the behaviour of the estimators of μ we need to know what happens to narrow and wide estimators of the respectively two and three model parameters.

Consider the score functions for the wide model, evaluated at the null point $(\xi_0,\sigma_0,0)$. Letting γ carefully tend to zero in expressions for the three partial log-derivatives of f leads to

$$\begin{pmatrix} U(y) \\ V(y) \\ W(y) \end{pmatrix} = \begin{pmatrix} \partial\log f(y,\xi_0,\sigma_0,0)/\partial\xi \\ \partial\log f(y,\xi_0,\sigma_0,0)/\partial\sigma \\ \partial\log f(y,\xi_0,\sigma_0,0)/\partial\gamma \end{pmatrix} = \begin{pmatrix} z/\sigma_0 \\ (z^2-1)/\sigma_0 \\ \frac{1}{4}z^4 - \frac{1}{2}z^2 - \frac{1}{4} \end{pmatrix}, \tag{2.5}$$

where $z = (y-\xi_0)/\sigma_0$, cf. (2.2). We shall also need the accompanying 3×3 size information matrix, the covariance matrix of these three, as Y has the $f(y,\xi_0,\sigma_0,0)$ distribution, i.e. is simply $N(\xi_0,\sigma_0^2)$. One finds

$$J_{\rm wide} = {\rm VAR}_0 \begin{pmatrix} Z/\sigma_0 \\ (Z^2-1)/\sigma_0 \\ \frac14 Z^4 - \frac12 Z^2 - \frac14 \end{pmatrix} = \begin{pmatrix} 1/\sigma_0^2 & 0 & 0 \\ 0 & 2/\sigma_0^2 & 2/\sigma_0 \\ 0 & 2/\sigma_0 & 7/2 \end{pmatrix}.$$

Note that the upper left hand 2×2 block $J_{\rm narr} = {\rm diag}(1/\sigma_0^2, 2/\sigma_0^2)$ is the information matrix of the narrow model, evaluated at (ξ_0,σ_0). For future reference we note that

$$J_{\rm wide}^{-1} = \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \frac76\sigma_0^2 & -\frac23\sigma_0 \\ 0 & -\frac23\sigma_0 & \frac23 \end{pmatrix}, \quad J_{\rm narr}^{-1} = \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \sigma_0^2/2 \end{pmatrix}. \tag{2.6}$$

LEMMA. *Let $\bar U_n$ denote average of $U(Y_i)$'s, and similarly for $\bar V_n$ and $\bar W_n$. Under the sequence of models P_n of (2.4),*

$$\begin{pmatrix} \sqrt n\,\bar U_n \\ \sqrt n\,\bar V_n \\ \sqrt n\,\bar W_n \end{pmatrix} \to_d \begin{pmatrix} 0+K \\ 2/\sigma_0\,\delta + L \\ 7/2\,\delta + M \end{pmatrix}$$

as $n\to\infty$, where $(K,L,M)' \sim \mathcal{N}_3\{0, J_{\rm wide}\}$.

PROOF: This follows essentially from the triangular version of the Lindeberg theorem. A key observation is that

$$f_n(y) \doteq f(y,\xi_0,\sigma_0,0)\{1+W(y)\delta/\sqrt n\}.$$

This implies that $\big(U(Y_i),V(Y_i),W(Y_i)\big)'$ has expected value $\big(0, \frac{2}{\sigma_0}\delta/\sqrt n, \frac72\delta/\sqrt n\big)$ plus $O(n^{-1})$ terms, and that its variance matrix is $J_{\rm wide} + O(\delta/\sqrt n)$. See also Section 2 of Hjort (1991). □

PROPOSITION 1. *Under model P_n of (2.4) one has*

$$\begin{pmatrix} \sqrt n(\widehat\xi_{\rm narr} - \xi_0) \\ \sqrt n(\widehat\sigma_{\rm narr} - \sigma_0) \end{pmatrix} \doteq_d J_{\rm narr}^{-1} \begin{pmatrix} \sqrt n\,\bar U_n \\ \sqrt n\,\bar V_n \end{pmatrix} \to_d \begin{pmatrix} 0+\sigma_0^2 K \\ \sigma_0\delta + \frac12\sigma_0^2 L \end{pmatrix},$$

in which $A_n \doteq_d B_n$ means that $A_n - B_n$ tends to zero in probability.

PROOF: This is essentially the familiar Taylor expansion argument, carried out in the present local neighbourhood framework. Note the bias term $(0,\sigma_0\delta)'$. The details are contained in more general arguments given in Section 2 of Hjort (1991). More direct methods of proof could also have been used since $\widehat\xi_{\rm narr}$ and $\widehat\sigma_{\rm narr}$ are relatively easy to work with. □

The wide method case is much more complicated because of the corner problem. Introduce

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = J_{\rm wide}^{-1} \begin{pmatrix} K \\ L \\ M \end{pmatrix} \sim \mathcal{N}_3\{0, J_{\rm wide}^{-1}\}. \tag{2.7}$$

Note that $K = (1/\sigma_0^2)A$, $L = (2/\sigma_0^2)B + (2/\sigma_0)C$. The limit in Proposition 1, written in terms of $(A, B, C)'$, becomes

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi}_{\rm narr} - \xi_0) \\ \sqrt{n}(\widehat{\sigma}_{\rm narr} - \sigma_0) \end{pmatrix} \to_d \begin{pmatrix} 0 + A \\ B + \sigma_0(C + \delta) \end{pmatrix} \sim \mathcal{N}_2\{\begin{pmatrix} 0 \\ \sigma_0\delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \frac{1}{2}\sigma_0^2 \end{pmatrix}\}. \tag{2.8}$$

Since I want my reader to join me for the main story I defer the proof of the following proposition to the appendix.

PROPOSITION 2. *Under model P_n of (2.4) one has*

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi} - \xi_0) \\ \sqrt{n}(\widehat{\sigma} - \sigma_0) \\ \sqrt{n}(\widehat{\gamma} - \delta/\sqrt{n}) \end{pmatrix} \to_d \begin{cases} \begin{pmatrix} A \\ B \\ C \end{pmatrix} & \text{if } C \geq -\delta, \\ \begin{pmatrix} A \\ B + \sigma_0(C + \delta) \\ -\delta \end{pmatrix} & \text{if } C \leq -\delta. \end{cases}$$

3. Calculating the tolerance distance. Our program is to use the delta method of linearisation in conjunction with Propositions 1 and 2 to reach limit distribution results for the narrow and wide estimators, and then to compute and compare mean squared errors.

First consider the narrow method. Using Proposition 1 we find

$$\begin{aligned} &\sqrt{n}\{\mu(\widehat{\xi}_{\rm narr}, \widehat{\sigma}_{\rm narr}, 0) - \mu(\xi_0, \sigma_0, \delta/\sqrt{n})\} \\ &\quad = \sqrt{n}\{\mu(\widehat{\xi}_{\rm narr}, \widehat{\sigma}_{\rm narr}, 0) - \mu(\xi_0, \sigma_0, 0)\} - \sqrt{n}\{\mu(\xi_0, \sigma_0, \delta/\sqrt{n}) - \mu(\xi_0, \sigma_0, 0)\} \\ &\quad \doteq_d \tfrac{\partial\mu}{\partial\xi}\sqrt{n}(\widehat{\xi}_{\rm narr} - \xi_0) + \tfrac{\partial\mu}{\partial\sigma}\sqrt{n}(\widehat{\sigma}_{\rm narr} - \sigma_0) - \sqrt{n}\tfrac{\partial\mu}{\partial\gamma}\delta/\sqrt{n} \\ &\quad \to_d \Lambda_{\rm narr} = \tfrac{\partial\mu}{\partial\xi}A + \tfrac{\partial\mu}{\partial\sigma}(B + \sigma_0(C + \delta)) - \tfrac{\partial\mu}{\partial\gamma}\delta, \end{aligned}$$

where the partial derivatives are computed at the null model $(\xi_0, \sigma_0, 0)$. The limit variable is normal with

$$\begin{aligned} E\Lambda_{\rm narr} &= b\delta = (\sigma_0\tfrac{\partial\mu}{\partial\sigma} - \tfrac{\partial\mu}{\partial\gamma})\,\delta, \\ \text{Var}\,\Lambda_{\rm narr} &= \tau_0^2 = \{(\tfrac{\partial\mu}{\partial\xi})^2 + \tfrac{1}{2}(\tfrac{\partial\mu}{\partial\sigma})^2\}\,\sigma_0^2. \end{aligned} \tag{3.1}$$

In particular the narrow method has risk $E\Lambda_{\rm narr}^2 = b^2\delta^2 + \tau_0^2$. See 6B for some consequences of this. This 'narrow result' is really contained in general results of Hjort (1991).

Next consider the wide method. Using Proposition 2 one finds

$$\begin{aligned} &\sqrt{n}\{\mu(\widehat{\xi}, \widehat{\sigma}, \widehat{\gamma}) - \mu(\xi_0, \sigma_0, \delta/\sqrt{n})\} \\ &\quad \doteq_d \tfrac{\partial\mu}{\partial\xi}\sqrt{n}(\widehat{\xi} - \xi_0) + \tfrac{\partial\mu}{\partial\sigma}\sqrt{n}(\widehat{\sigma} - \sigma_0) + \{(\tfrac{\partial\mu}{\partial\gamma}) + O(1/\sqrt{n})\}\sqrt{n}(\widehat{\gamma} - \delta/\sqrt{n})) \\ &\quad \to_d \Lambda_{\rm wide} = \begin{cases} \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}B + \frac{\partial\mu}{\partial\gamma}C & \text{if } C \geq -\delta, \\ \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}(B + \sigma_0(C + \delta)) - \frac{\partial\mu}{\partial\gamma}\delta & \text{if } C \leq -\delta. \end{cases} \end{aligned}$$

This is not a normal distribution. We calculate its mean squared error by conditioning on the value of C. Using (2.6) and (2.7) and ordinary techniques one finds

$$\begin{pmatrix} A \\ B \end{pmatrix} | \{C = c\} \sim \mathcal{N}_2\{\begin{pmatrix} 0 \\ -\sigma_0 c \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \frac{1}{2}\sigma_0^2 \end{pmatrix}\}. \tag{3.2}$$

Further calculations show that

$$\Lambda_{\rm wide}|\{C=c\} \sim \begin{cases} \mathcal{N}\{-bc,\tau_0^2\} & \text{if } c \geq -\delta, \\ \mathcal{N}\{b\delta,\tau_0^2\} & \text{if } c \leq -\delta. \end{cases}$$

Accordingly $E\{\Lambda_{\rm wide}^2|C=c\}$ is $b^2c^2+\tau_0^2$ if $c \geq -\delta$ and $b^2\delta^2+\tau_0^2$ if $c \leq -\delta$, and

$$E\Lambda_{\rm wide}^2 = b^2\,E\big[C^2I\{C \leq -\delta\} + \delta^2 I\{C \leq -\delta\}\big] + \tau_0^2.$$

We are now in a position to find out when the narrow and risky estimator is better than the wide and safe one. From (2.7) we can write $C=\kappa N$ where $\kappa^2=\frac{2}{3}$ and N is normal $(0,1)$. From (3.1), assuming $b \neq 0$, it is clear that the narrow method is better than the wide one if and only if

$$\delta^2 \leq E\big[\kappa^2N^2I\{N \geq -\delta/\kappa\} + \delta^2 I\{N \leq -\delta/\kappa\}\big],$$

or

$$a^2 \leq E\big[N^2I\{T \geq -a\} + a^2I\{N \leq -a\}\big] = \Phi(a) - a\phi(a) + a^2(1-\Phi(a)), \tag{3.3}$$

using $a=\delta/\kappa$. But this is equivalent to $0 \leq a \leq 0.8399$, as borne out by numerical computations. This means $0 \leq \delta \leq 0.8399\sqrt{2/3} = 0.6858$, and we have reached

RESULT. *(i) The case where* $b = \sigma_0\frac{\partial\mu}{\partial\sigma} - \frac{\partial\mu}{\partial\gamma} = 0$ *is rather trivial; this typically corresponds to a parameter estimand* μ *functionally independent of* σ *and* γ *at* $\gamma=0$. *In this case* $\widehat{\mu}_{\rm wide}$ *and* $\widehat{\mu}_{\rm narr}$ *are asymptotically equivalent, regardless of* δ. *(ii) In the more interesting case* $b \neq 0$, *the narrow model based estimator is better than or as good as the wider model based estimator if and only if* $0 \leq \delta \leq 0.6858$, *or* $0 \leq \gamma \leq 0.6858/\sqrt{n}$, *or degrees of freedom* $m \geq \sqrt{n}/0.6858 = 1.4582\sqrt{n}$.

4. A fuller story: compromise estimators. The two estimators (2.3) that have been considered so far have both somewhat extreme attitudes. The first is a firm believer and the second a firm disbeliever in normality. This section looks at some comprimising methods that are designed to work well both under 'close to normal' and 'distinctively nonnormal' conditions. See also Section 5 in Hjort (1991).

We have shown in Proposition 2 that $\sqrt{n}(\widehat{\gamma}-\delta/\sqrt{n})$ tends to $\max\{C,-\delta\}$ in distribution, where $C \sim \mathcal{N}\{0,\kappa^2\}$ and $\kappa^2=\frac{2}{3}$. Now shift attention to $T_n=\sqrt{n}\widehat{\gamma}/\kappa$, the natural statistic for testing $\gamma=0$ (normality) against $\gamma>0$ (t-ness). Using $a=\delta/\kappa$ again we have

$$T_n \to_d a + \kappa^{-1}\max\{C,-\delta\} = T \vee 0, \quad \text{where } T = a + C/\kappa \sim \mathcal{N}\{a,1\}, \tag{4.1}$$

and

$$\begin{aligned} \Lambda_{\rm narr} &= b\delta + \tfrac{\partial\mu}{\partial\xi}A + \tfrac{\partial\mu}{\partial\sigma}(B+\sigma_0\kappa(T-a)), \\ \Lambda_{\rm wide} &= \begin{cases} \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}B + \frac{\partial\mu}{\partial\gamma}\kappa(T-a) & \text{if } T>0, \\ \frac{\partial\mu}{\partial\xi}A + \frac{\partial\mu}{\partial\sigma}(B+\sigma_0\kappa T) - \frac{\partial\mu}{\partial\gamma}a\kappa & \text{if } T \leq 0. \end{cases} \end{aligned} \tag{4.2}$$

Study the general estimator

$$\mu^* = \{1 - w(T_n)\}\,\widehat{\mu}_{\text{narr}} + w(T_n)\,\widehat{\mu}_{\text{wide}}, \tag{4.3}$$

where $w(T_n)$ is some appropriate weight function, assumed only to be continuous at zero (where the limit $T \vee 0$ of T_n has positive probability) and almost everywhere on $(0, \infty)$. Then, by the continuous mapping theorem,

$$\sqrt{n}(\mu^* - \mu_{\text{true}}) \to \Lambda = \{1 - w(T \vee 0)\}\,\Lambda_{\text{narr}} + w(T \vee 0)\,\Lambda_{\text{wide}}. \tag{4.4}$$

Diligent work shows that

$$\Lambda | \{T = t\} \sim \begin{cases} \mathcal{N}\{b\kappa(a - w(t)t), \tau_0^2\} & \text{if } t > 0, \\ \mathcal{N}\{b\kappa a, \tau_0^2\} & \text{if } t \le 0, \end{cases}$$

from which it follows that the limit risk for the (4.3) estimator can be written

$$E\Lambda^2 = \tfrac{2}{3}b^2 R(a) + \tau_0^2, \quad \text{where } R(a) = E_a\big[(w(T)T - a)^2 I\{T > 0\} + a^2 I\{T \le 0\}\big]. \tag{4.5}$$

Observe that $R(a)$ is the risk function, under squared error loss, for the estimator $\widehat{a}(T \vee 0) = w(T \vee 0)(T \vee 0)$ for a nonnegative parameter a, based on observing the single variable $T \vee 0$, where $T \sim \mathcal{N}\{a, 1\}$. There is accordingly a one-to-one correspondence between estimators μ^* of type (4.3) for a general $\mu(\xi, \sigma, \gamma)$ and estimators $a^*(t) = w(t)t$ for a in the structurally very simple one-observation problem. The behaviour of any given μ^* can be studied quite simply in terms of its associated $R(a)$ function, and any reasonable a-estimator method can be transported to a reasonable μ-estimator, via $w(t) = a^*(t)/t$.

What are interesting values of a? We have $a = \delta/\kappa$ and $m = 1/\gamma = \sqrt{n}/\delta = \sqrt{1.5\,n}/a$, and T_n of (4.1) detects non-normality ($m < \infty$) with probability $\Phi(a - 1.645)$ (using level 5%). This means that T_n detects a-values beyond 4 with probability at least 0.99. We may think of a-values beyond 4, or $m \le 0.306\sqrt{n}$, as being t-departures from normality that should be clearly visible from data. This tentatively suggests that estimators of the type (4.3) should be used with $w(t)$ close to 1 for $t \ge 4$ and with small risk behaviour for $R(a)$ when $a \le 4$.

There follows a briefly annotated list of interesting choices for $w(T_n)$ in (4.3).

(i) The narrow method uses $w(t) = 0$, and corresponds to using $\widehat{a}_{\text{narr}}(t) = 0$ to estimate a. Its risk is $R_{\text{narr}}(a) = a^2$, which is good for a small (m large) but disastrous for a large (m small).

(ii) The wide method has $w(t) = 1$, and corresponds to $\widehat{a}_{\text{wide}}(t) = t \vee 0$ to estimate a. Its risk is

$$R_{\text{wide}}(a) = E_a\big[(T - a)^2 I\{T \ge 0\} + a^2 I\{T \le 0\}\big] = \Phi(a) - a\phi(a) + a^2(1 - \Phi(a)),$$

cf. (3.3). It starts at 0.50 at zero and climbs towards 1. This estimator is minimax. Its risk is above .99 for $a \ge 2.67$. Again: if $0 \le 0.8399$ then the narrow is best and if $a > 0.8399$ then the wide is best.

(iii) Try out $w(t) = w$, a constant. We may compute the resulting $R(a)$, and minimise w.r.t. the choice of w. The best choice, expressed in terms of the parameter point a, is

$$w_0(a) = \frac{a^2\Phi(a) + a\phi(a)}{(a^2+1)\Phi(a) + a\phi(a)}.$$

A simple idea is then to insert T_n for a, i.e. using $\widehat{a}_{\rm ratio} = w_0(t)t$ to estimate a and (4.3) with $w_0(T_n)$ to estimate μ. $R_{\rm ratio}(a)$ starts at 0.249 and is better than $R_{\rm wide}(a)$ for $a \leq 1.32$, and is never much worse. Its maximum is 1.223, at $a = 2.90$, after which it decreases towards 1. The narrow is better than the present one only for $a \leq 0.68$, and quickly becomes much worse after that.

(iv) Some natural Bayesian/empirical Bayesian ideas are as follows. Assume a is distributed like $|N(0,\tau^2)|$, i.e. with prior distribution $\pi(a) = \frac{2}{\tau}\phi(a/\tau)$ on $[0,\infty)$. The Bayes solution associated with the loss function implicit in (4.5) can be seen to be the familiar $E\{a|T=t\}$ if $t > 0$ and an arbitrary value can be used when $T \vee 0 = 0$, i.e. when information on T is $T \leq 0$. In the present case the Bayes solution becomes

$$\widehat{a}_\tau(t) = E\{a|t\} = \nu t + \sqrt{\nu}\phi(\sqrt{\nu}t)/\Phi(\sqrt{\nu}t), \quad \text{where } \nu = \tau^2/(\tau^2+1).$$

Since $E_a T = a$ and $Ea^2 = \tau^2$ a simple empirical estimate for ν is $T^2/(T^2+1)$. This leads to

$$\widehat{a}_{\rm eb}(t) = \frac{t^2}{t^2+1}t + \frac{t}{\sqrt{t^2+1}}\phi\Big(\frac{t}{\sqrt{t^2+1}}t\Big)\Big/\Phi\Big(\frac{t}{\sqrt{t^2+1}}t\Big).$$

Performance: $R_{\rm eb}(a)$ starts at 0.337 and is better than $R_{\rm wide}(a)$ for $a \leq 2.09$, and is never much worse. It is not quite as good as the narrow method when $a \leq 0.67$, but quickly becomes much better after that. It reaches its maximum value of only 1.147 at $a = 3.75$, and decreases towards 1 thereafter.

(v) The limit of the Bayes rules above, when $\tau \to \infty$, is $\widehat{a}_{\rm vag}(t) = t + \phi(t)/\Phi(t)$. This is also the Bayes solution under a vague flat prior on the halfline. It is minimax like the $\widehat{a}_{\rm wide}$, but has a differently shaped risk function, see Figure 1.

(vi) Finally we could mention pre-test and related estimators. The if-else of pre-test estimator uses $w(t) = 0$ if $t \leq d$ and $w(t) = 1$ if $t > d$ in (4.3), and corresponds to $\widehat{a}_{\rm pre}(t) = 0$ if $t \leq d$ and $\widehat{a}_{\rm pre}(t) = t$ if $t > d$, for suitable cut-off value d. The theory of Section 3 could invite $d = 0.8399$, for example. It has risk

$$R_{\rm pre}(a) = \Phi(a-d) + a^2\{1-\Phi(a-d)\} - (a-d)\phi(d-a).$$

A related but smoother version is the limited translation variety $\widehat{a}_{\rm lim}(t) = 0$ if $t \leq d$ and $\widehat{a}_{\rm lim}(t) = t - d$ if $t > d$. This corresponds to using $w(t) = 0$ if $t \leq d$ and $w(t) = 1 - d/t$ if $t > d$. The risk function becomes

$$R_{\rm lim}(a) = (1+d^2)\Phi(a-d) + a^2\{1-\Phi(a-d)\} - (a+d)\phi(a-d),$$

with maximum value $1 + d^2$ occurring at infinity.

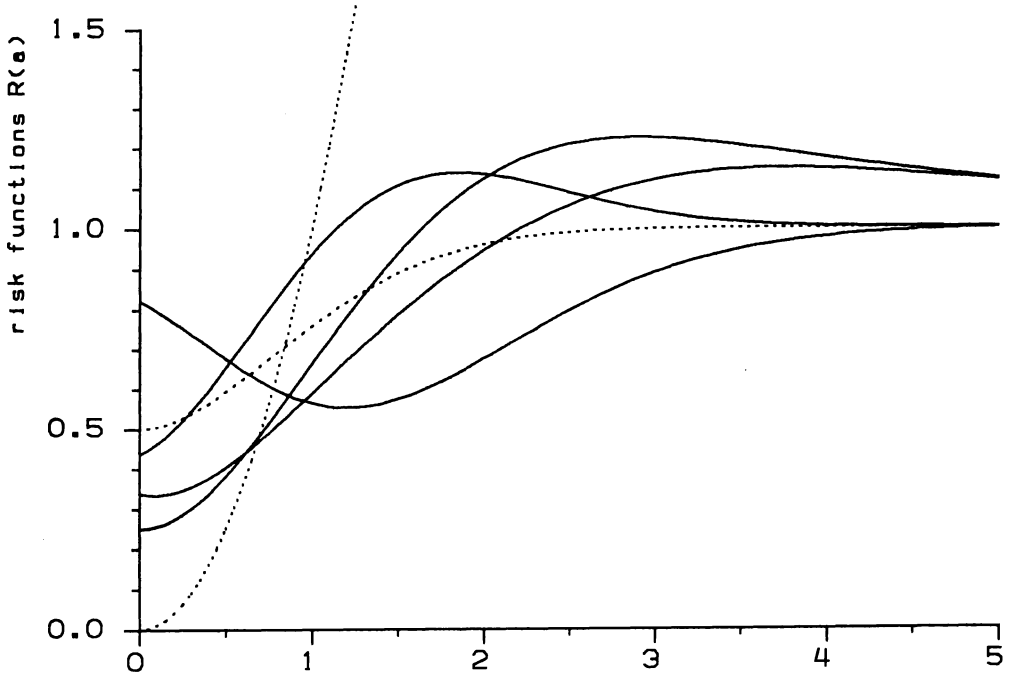

FIGURE 1. *Risk functions $R(a)$ are shown for six procedures, corresponding to six choices of weight function $w(.)$ in (4.3). Risks for the wide and the narrow methods start at respectively .500 and .000 and are shown with dotted lines. The other four are the ratio method (starting at .249), the empirical Bayes method (starting at .337), the pre-test method with $d = .8399$ (starting at .436), and Bayes solution under uniform prior on the halfline (starting at .819).*

It is worth mentioning that in the general but regular case treated in Hjort (1991), where $a = \delta/\kappa$ can vary freely on the line, then methods (v) and (ii) above become equivalent, as do ideas (iii) and (iv).

Several risk functions $R(a)$ are plotted in Figure 1. [Tables and computer programs for these and several other risk functions are available from the author upon corteous request.] Overall both the ratio method (iii) and the empirical Bayes method (iv) seem to be quite satisfactory solutions: they are almost as good as the wide method under distinctively non-normal conditions and are considerably better than the wide method under close-to-normal conditions.

5. Extension to regression models. Suppose

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \sigma Z_i = x_i'\beta + \sigma Z_i, \tag{5.1}$$

where $\beta_1, \ldots, \beta_p$ are regression parameters and the Z_i's are i.i.d. from a t_m-distribution. How risky are the ordinary methods, that all have $m = \infty$ as underlying assumption? How statistically noisy are the more ambitious $p+2$-parameter methods that employ ML estimates for $\beta_1, \ldots, \beta_p, \sigma, m$? For example, should one use

$$\widehat{\mu}_{\text{narr}} = \sqrt{2/\pi}\,\widehat{\sigma}_{\text{narr}} \quad \text{or} \quad \widehat{\mu}_{\text{wide}} = \sqrt{1/\pi}\,\{\Gamma(\tfrac{1}{2}\widehat{m}_{\text{wide}} - \tfrac{1}{2})/\Gamma(\tfrac{1}{2}\widehat{m}_{\text{wide}})\}\,\widehat{\sigma}_{\text{wide}}$$

to estimate $\mu = E|Y(x) - x'\beta|$, the expected distance from regression curve to data point?

Let us briefly indicate how results from earlier sections extend to this situation. Let β_0 and σ_0 be arbitrary but fixed, and let $\gamma = 1/m$ tend to zero like $\gamma = \gamma_n = \delta/\sqrt{n}$. The score function becomes

$$\begin{pmatrix} U(y_i) \\ V(y_i) \\ W(y_i) \end{pmatrix} = \begin{pmatrix} \partial \log f(y_i, \beta_0, \sigma_0, 0)/\partial\beta \\ \partial \log f(y_i, \beta_0, \sigma_0, 0)/\partial\sigma \\ \partial \log f(y_i, \beta_0, \sigma_0, 0)/\partial\gamma \end{pmatrix} = \begin{pmatrix} z_i x_i/\sigma_0 \\ (z_i^2 - 1)/\sigma_0 \\ \frac{1}{4}z_i^4 - \frac{1}{2}z_i^2 - \frac{1}{4} \end{pmatrix},$$

in which $z_i = (y_i - x_i'\beta_0)/\sigma_0$. The $(p+2) \times (p+2)$ information matrix becomes

$$J_{\text{wide}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \text{VAR}_0 \begin{pmatrix} Z_i x_i/\sigma_0 \\ (Z_i^2 - 1)/\sigma_0 \\ \frac{1}{4}Z_i^4 - \frac{1}{2}Z_i^2 - \frac{1}{4} \end{pmatrix} = \begin{pmatrix} D/\sigma_0^2 & 0 & 0 \\ 0 & 2/\sigma_0^2 & 2/\sigma_0 \\ 0 & 2/\sigma_0 & 7/2 \end{pmatrix},$$

in which it is assumed that the $p \times p$ design matrix D, the limit in probability of $D_n = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$, exists. We note that

$$J_{\text{wide}}^{-1} = \begin{pmatrix} \sigma_0^2 D^{-1} & 0 & 0 \\ 0 & \frac{7}{6}\sigma_0^2 & -\frac{2}{3}\sigma_0 \\ 0 & -\frac{2}{3}\sigma_0 & \frac{2}{3} \end{pmatrix}, \quad J_{\text{narr}}^{-1} = \begin{pmatrix} \sigma_0^2 D^{-1} & 0 \\ 0 & \sigma_0^2/2 \end{pmatrix}.$$

The parallel to Section 2's Lemma is that

$$\begin{pmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_n \\ \sqrt{n}\,\bar{W}_n \end{pmatrix} \to_d \begin{pmatrix} 0 + K \\ 2/\sigma_0\,\delta + L \\ 7/2\,\delta + M \end{pmatrix},$$

where $(K, L, M)' \sim \mathcal{N}_{p+2}\{0, J_{\text{wide}}\}$. This is true by the triangular Lindeberg theorem under the familiar condition $\frac{1}{n}\max_{i\le j\le n}(x_{i,j} - \bar{x}_i)^2 \to 0$ for each i. For the familiar normality-based (least-squares-type) estimators one finds

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\widehat{\beta}_{\text{narr}} - \beta_0) \\ \sqrt{n}(\widehat{\sigma}_{\text{narr}} - \sigma_0) \end{pmatrix} &\doteq_d J_{\text{narr}}^{-1} \begin{pmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_n \end{pmatrix} \to_d \begin{pmatrix} 0 + \sigma_0^2 K \\ \sigma_0\delta + \frac{1}{2}\sigma_0^2 L \end{pmatrix} \\ &= \begin{pmatrix} 0 + DA \\ B + \sigma_0(C + \delta) \end{pmatrix} \sim \mathcal{N}_{p+1}\{\begin{pmatrix} 0 \\ \sigma_0\delta \end{pmatrix}, \begin{pmatrix} \sigma_0^2 D^{-1} & 0 \\ 0 & \frac{1}{2}\sigma_0^2 \end{pmatrix}\}. \end{aligned}$$

writing $(A, B, C)'$ for $J_{\text{wide}}^{-1}(K, L, M)'$, which is $\mathcal{N}_{p+2}\{0, J_{\text{wide}}^{-1}\}$. Next, regarding the ML estimators $\widehat{\beta}$, $\widehat{\sigma}$, $\widehat{\gamma}$ in the wider $p+2$-parameter model, Proposition 2 with proof can be lifted mutatis mutandis and becomes

$$\begin{pmatrix} \sqrt{n}(\widehat{\beta} - \beta_0) \\ \sqrt{n}(\widehat{\sigma} - \sigma_0) \\ \sqrt{n}\widehat{\gamma}/\kappa \end{pmatrix} \to_d \begin{cases} \begin{pmatrix} A \\ B \\ T \end{pmatrix} & \text{if } T \ge 0, \\ \begin{pmatrix} A \\ B + \sigma_0\kappa T \\ 0 \end{pmatrix} & \text{if } T \le 0. \end{cases}$$

The rest of the story is very similar to that of Sections 3 and 4. The limit variables Λ_{narr} and Λ_{wide} are as in (4.2), only with $(\frac{\partial\mu}{\partial\beta})'DA$ replacing $\frac{\partial\mu}{\partial\xi}A$, and A is now

$\mathcal{N}_p\{0,\sigma_0^2 D^{-1}\}$ and not merely $\mathcal{N}\{0,\sigma_0^2\}$. Conditioning on T one finds in the end that the main result (4.5) is true, with

$$b = \sigma_0 \tfrac{\partial\mu}{\partial\sigma} - \tfrac{\partial\mu}{\partial\gamma}, \quad \tau_0^2 = \{(\tfrac{\partial\mu}{\partial\beta})' D^{-1} (\tfrac{\partial\mu}{\partial\beta}) + \tfrac{1}{2}(\tfrac{\partial\mu}{\partial\sigma})^2\}\,\sigma_0^2. \tag{5.2}$$

Section 3's main result about $m \geq 1.458\sqrt{n}$ is also true verbatim. And for the problem of performing linear regression analysis when the residuals could have fatter tails than the normal, a natural poposal is to use

$$\mu^* = \{1 - w(\sqrt{1.5\,n}\,\widehat{\gamma})\}\,\mu(\widehat{\beta}_{\rm narr}, \widehat{\sigma}_{\rm narr}, 0) + w(\sqrt{1.5\,n}\,\widehat{\gamma})\,\mu(\widehat{\beta},\widehat{\sigma},\widehat{\gamma}), \tag{5.3}$$

where $w(.)$ is as in (iii) or (iv) of Section 4.

6. Additional remarks.

6A. Some estimands. To illustrate both the general formulae and the relative importance of bias and estimation noise, let us go through a short list of important estimands.

(i) Let $\mu = x'\beta$, the regression curve at a specific point. Then $b = 0$ and all compromise estimators become asymptotically equivalent, with $\tau_0^2 = \frac{1}{2} x' D^{-1} x\,\sigma_0^2$ as limiting normalised risk. Thus familiar least-squares estimates are sufficiently precise even in the presence of t-ness, and the same is true in other cases where the estimand only depends upon $\beta_1,\ldots,\beta_p$. (ii) Let $\mu = E|Y(x) - x'\beta|$, our starting example of Section 5. Then $\mu = \sigma\,E|Z|$, where Z is t_m-distributed, and clever calculations show that $b = \frac{1}{2}\sigma_0\phi(0)$, $\tau_0^2 = 2\phi(0)^2\sigma_0^2$. This gives

$$\text{risk} = \frac{\sigma_0^2}{\pi}\Big\{\frac{1}{12}R(a) + 1\Big\}$$

for the limit distribution version of n times mean squared error for μ^*, see (4.5). (iii) Let μ be the p-th quantile of the distribution for $Y(x)$ at x. It is for example often useful and illuminating to draw the nine regression deciles (corresponding to $p = j/10$) in the same diagram, as functions of x. Then $\mu = x'\beta + \sigma G^{-1}(p,m)$ in the notation of Section 1. One can work out a suitable expression for $\partial\mu/\partial\gamma$, and then find b and τ_0 of (5.2). The end result is

$$\text{risk} = \big[\tfrac{2}{3}\{z_p + A(z_p)/\phi(z_p)\}^2 R(a) + x' D^{-1} x + \tfrac{1}{2} z_p^2\big]\,\sigma_0^2,$$

in which $z_p = \Phi^{-1}(p)$ and $A(t) = \int_{-\infty}^t \phi(z) W(z)\,dz$, and $W(z)$ is as in (2.5). (iv) The case of a probability $\mu = \Pr\{Y(x) \leq y\} = G\big((y - x'\beta)/\sigma, m\big)$ is similar to but simpler than case (iii). The same expression for risk emerges, with $z(y) = (y - x'\beta_0)/\sigma_0$ replacing z_p.

6B. False confidence. We proved in Section 3 that $\sqrt{n}(\widehat{\mu}_{\rm narr} - \mu_{\rm true})$ tends to $\mathcal{N}\{b\delta, \tau_0^2\}$ under the (2.4) sequence of models. Traditional normality-based inference uses in essence that the limit is $\mathcal{N}\{0,\tau_0^2\}$. Accordingly $b^2\delta^2$ is the invisible extra burden associated with using the normality-based estimator when in fact the wider model (2.4) is true. Consequences of this include that traditional normality-based confidence intervals and testing procedures behave incorrectly; the intervals have adequate length but are incorrectly placed, and the tests have too high significance levels. If $\text{CI}_{\rm narr} = \widehat{\mu} \pm 1.645\widehat{\tau}_0/\sqrt{n}$, for example, then the coverage probability converges to $\Pr\{|\mathcal{N}\{b\delta/\tau_0, 1\}| \leq 1.645\}$, which is strictly less than 90% unless $b = 0$ or $\delta = 0$. See also Section 4H of Hjort (1991).

6C. How far away is $1.458\sqrt{n}$*?* One can test normality ($\gamma = 0$) against t-ness ($\gamma > 0$) using $T_n = \sqrt{1.5\,n}\,\widehat{\gamma}$, see Section 4. The limit distribution under normality is $\max\{N, 0\}$ where N is standard normal. The $T_n > 1.645$ test has (asymptotic) level 5% and power $\Phi(a - 1.645)$. One way of quantifying the distance from normality to the first intolerable t-distribution is in terms of $\Phi(0.8399 - 1.645) = 0.210$, the probability of detecing this amout of t-ness. The corresponding detection probability figure is 0.329 for the case of a 10% level test.

Other distance measures are possible; see Hjort (1991, Sections 4B and 4C) for other proposals and interpretations. The L_1-distance $\int |f_\gamma - f_0|\,dy$ is approximately $0.434/\sqrt{n}$.

6D. A quasi-extension of the t-distribution with negative degrees of freedom. It was necessary to use non-standard corner asymptotics to reach results in Sections 3–5. The problems would have been much simpler to solve if the parameter space for $\gamma = 1/m$ had included zero as an inner point, i.e. if the model had permitted negative values of γ. This is not only a technical but also a statistical point, since data sets could easily display lighter-than-normal tails (negative kurtosis), and in a way it is an artificial facet of the smooth transition from t-ness to normality (letting $m \to \infty$) that it has stop right there.

It is therefore tempting to by-pass the whole t-model and create a new alternative model $f(y, \xi, \sigma, \gamma)$ that permits negative values of γ. Inspired by (2.2) one could try

$$f(y, \xi, \sigma, \gamma) = \phi\Big(\frac{y-\xi}{\sigma}\Big)\frac{1}{\sigma}\Big\{1 + \gamma A\Big(\frac{y-\xi}{\sigma}\Big)\Big\}$$

for suitable $A(z)$-function. Natural desiderata are (i) $A(z)$ is symmetric about zero; (ii) the model is defined for γ's in an interval around zero; (iii) the density decreases with y for $y \geq \xi$; (iv) the kurtosis is positive for $\gamma > 0$ and negative for $\gamma < 0$.

This can be achieved in various ways. Note that $A(z)$ has to be bounded because of (ii). Having a quasi-extension of the t-distribution in mind we try

$$A(z) = \begin{cases} \frac{1}{4}z^4 - \frac{1}{2}z^2 - a(c) & \text{if } |z| \leq c, \\ \frac{1}{4}c^4 - \frac{1}{2}c^2 - a(c) & \text{if } |z| \geq c, \end{cases}$$

where $a(c) = \frac{1}{4} - (\frac{1}{2}c^3 + \frac{1}{2}c)\phi(c) + (\frac{1}{2}c^4 - c^2 - \frac{1}{2})\{1 - \Phi(c)\}$ ensures the necessary $\int \phi(z)A(z)\,dz = 0$. (With some extra efforts the family could be smoothed at the cut-off points $\pm c$.) Judicious analysis shows that (ii)–(iv) hold for γ-values in a suitable $(l(c), r(c))$ interval around zero, at least when $c \geq \sqrt{6} = 2.4495$. I have found formulae for the various necessary quantities (J_{wide} etc.) in terms of c. If c is chosen large then $l(c)$ closes in on zero, so we might as well choose $c = \sqrt{6}$, for which the permissible interval is $(-0.171, 0.124)$. This defines a quasi-t-distribution with degrees of freedom m permitted to go from about 8 to infinity and over the top down to about -6. Figure 2 shows the quasi-t with 10 and -10 degrees of freedom. The quasi-t and the t are almost identical when $m \geq 10$ ($0 \leq \gamma \leq 0.10$). The kurtosis curve has derivative 1.244 at $\gamma = 0$ for this quasi-t family of probability densitities, and further analysis (but with no corner asymptotics required) shows that the normal model can tolerate deviation up to $|\gamma| \leq 1.895/\sqrt{n}$.

6E. Other problems with similar characteristics. There are other natural extensions of the basic normal model that also involve problems with corners of parameter spaces,

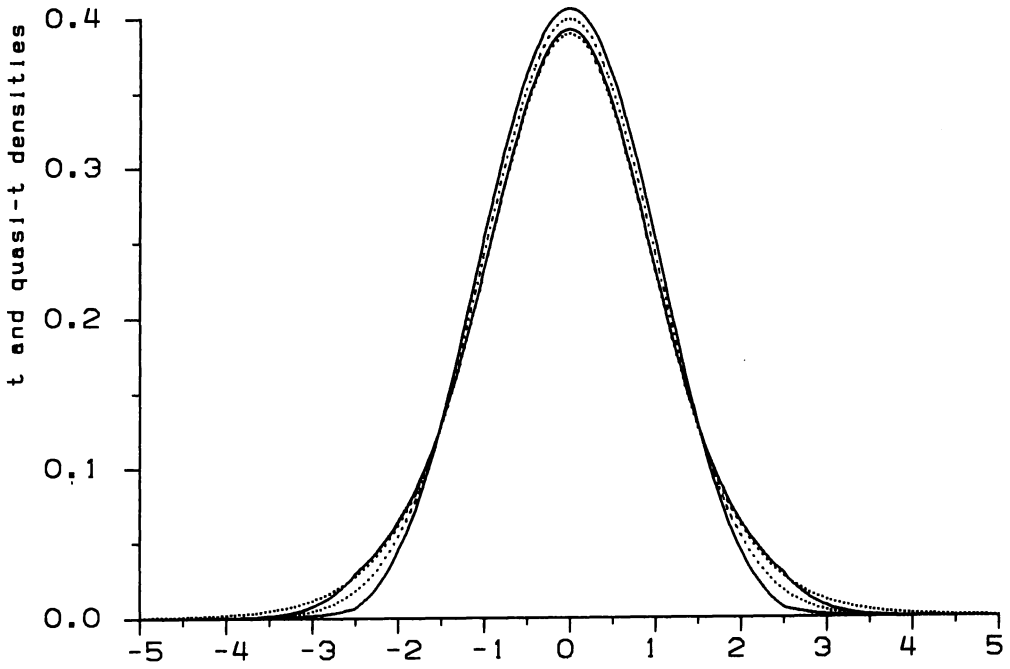

FIGURE 2. *Four probability densities are shown: The standard normal (t with $m = \infty$) and the t-density with ten degrees of freedom are shown with dotted lines. The quasi-t with respectively ten and minus ten degrees of freedom are shown with solid lines. The t_{10} and the quasi-t_{10} are quite close, and have positive kurtoses. The quasi-t_{-10} has negative kurtosis.*

and where methods of this paper can be useful. One example is the contamination model, where

$$f(y,\xi,\sigma,\varepsilon) = (1-\varepsilon)\mathcal{N}\{\xi,\sigma^2\}(y) + \varepsilon\mathcal{N}\{\xi,(k\sigma)^2\}(y)$$

for some known or unknown $k > 1$ and for some mixture parameter $\varepsilon \geq 0$. A natural question is how much contamination the normal model can tolerate.

Appendix: Proof of Proposition 2. The log-likelihood function can be written

$$\begin{aligned} L_n(\xi,\sigma,\gamma) = \sum_{i=1}^n \log f(y_i,\xi,\sigma,\gamma) &= -\log(2\pi)^{1/2} - \log\sigma - \tfrac{1}{2}\frac{1}{n}\sum_{i=1}^n (y_i-\xi)^2/\sigma^2 \\ &+ \gamma\sum_{i=1}^n R\big((y_i-\xi)/\sigma\big) - \tfrac{1}{2}\gamma^2\sum_{i=1}^n S\big((y_i-\xi)/\sigma\big) + O_p(n\gamma^3), \end{aligned} \tag{A.1}$$

by (2.2), in which

$$R(z) = \tfrac{1}{4}z^4 - \tfrac{1}{2}z^2 - \tfrac{1}{4} \quad \text{and} \quad S(z) = -\tfrac{1}{2}z^4 + \tfrac{1}{3}z^6.$$

The limit in probability of $\frac{1}{n}L_n(\xi,\sigma,\delta/\sqrt{n})$, under sequence (2.4), is seen to be $-\log(2\pi)^{1/2} - \log\sigma - \frac{1}{2}\{(\xi-\xi_0)^2 + \sigma_0^2\}/\sigma^2$, uniformly over compact sets. It follows that the sequence of ML estimators $\widehat{\xi}$, $\widehat{\sigma}$ must converge in probability to the parameter values that maximise

this limit, i.e. to the underlying ξ_0, σ_0. By working with the γ-related part of (A.1) one can similarly show that $\widehat{\gamma} = \widehat{\delta}/\sqrt{n}$ must converge to zero in probability.

Let in what follows $I_n(\xi,\sigma,\gamma)$ be the 3×3 matrix with elements $\sum_{i=1}^n (\partial^2/\partial\xi^2)\log f(y_i, \xi,\sigma,\gamma)$ &cetera. If $\widetilde{\xi}_n$, $\widetilde{\sigma}_n$, $\widetilde{\gamma}_n$ tend to respectively ξ_0, σ_0, 0 in probability, still under the (2.4) sequence of models, then $-\frac{1}{n}I_n(\widetilde{\xi}_n,\widetilde{\sigma}_n,\widetilde{\gamma}_n) \to_p J_{\rm wide}$. This holds since direct inspection shows

$$-\frac{1}{n}I_n(\widetilde{\xi}_n,\widetilde{\sigma}_n,\widetilde{\gamma}_n) = -\frac{1}{n}I_n(\xi_0,\sigma_0,0) + O_p(|\widetilde{\xi}_n-\xi_0| + |\widetilde{\sigma}_n-\sigma_0| + \widetilde{\gamma}_n),$$

and the first term here can be shown to converge to $J_{\rm wide}$, under (2.4), using ordinary methods.

There are two possibilities regarding the maximisers of (A.1). Either data $(y_1,\ldots,y_n)$ are such that maximum occurs for some $\widehat{\gamma} > 0$, or it occurs for $\widehat{\gamma} = 0$. In the first case the ML values are solutions to $\partial L_n/\partial\xi = 0$, $\partial L_n/\partial\sigma = 0$, $\partial L_n/\partial\gamma = 0$, and the familiar Taylor argument yields

$$\begin{pmatrix} \sqrt{n}(\widehat{\xi}-\xi_0) \\ \sqrt{n}(\widehat{\sigma}-\sigma_0) \\ \sqrt{n}(\widehat{\gamma}-0) \end{pmatrix} = \{-I_n(\widetilde{\xi},\widetilde{\sigma},\widetilde{\gamma})/n\}^{-1} \begin{pmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_n \\ \sqrt{n}\,\bar{W}_n \end{pmatrix}$$

for suitable $(\widetilde{\xi},\widetilde{\sigma},\widetilde{\gamma})$ somewhere between $(\xi_0,\sigma_0,0)$ and $(\widehat{\xi},\widehat{\sigma},\widehat{\gamma})$, see the definition in (2.5). In the second case $L_n(\xi,\sigma,\gamma)$ decreases in $\gamma \geq 0$, and the ML estimators are $(\widehat{\xi}_{\rm narr},\widehat{\sigma}_{\rm narr},0)$. Let Ω_n be the set of $(y_1,\ldots,y_n)$ for which the first case happens. Then $\sqrt{n}(\widehat{\xi}-\xi_0,\widehat{\sigma}-\sigma_0,\widehat{\gamma}-0)'$ becomes

$$J_{\rm wide}^{-1}\begin{pmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_n \\ \sqrt{n}\,\bar{W}_n \end{pmatrix} + O_p(n^{-1/2}) \quad \text{or} \quad \begin{pmatrix} J_{\rm narr}^{-1}\begin{pmatrix} \sqrt{n}\,\bar{U}_n \\ \sqrt{n}\,\bar{V}_n \end{pmatrix} + O_p(n^{-1/2}) \\ 0 \end{pmatrix}$$

as respectively Ω_n is in command or not. It turns out that Ω_n happens or not according to whether

$$\Delta_n = -\tfrac{2}{3}\sigma_0\sqrt{n}\,\bar{V}_n + \tfrac{2}{3}\sqrt{n}\,\bar{W}_n > 0 \text{ or } \leq 0, \tag{A.2}$$

to a first order approximation. Δ_n is the third component of $J_{\rm wide}^{-1}\sqrt{n}(\bar{U}_n,\bar{V}_n,\bar{W}_n)'$, and the precise statement is that $I(\Omega_n) - I\{\Delta_n > 0\}$ goes to zero in probability under the (2.4) regime of models. Using this result, the Lemma, and (2.7)–(2.8) in tandem yields the statement of Proposition 2, by the continuous mapping theorem on $\sqrt{n}(\bar{U}_n,\bar{V}_n,\bar{W}_n)'$.

To prove that Ω_n and $\{\Delta_n > 0\}$ are asymptotically equivalent events, consider once more the second half of (A.1), which is

$$\delta\frac{1}{\sqrt{n}}\sum_{i=1}^n R\big((y_i-\xi)/\sigma\big) - \tfrac{1}{2}\delta^2\frac{1}{n}\sum_{i=1}^n S\big((y_i-\xi)/\delta\big) + O_p(\delta^3/\sqrt{n}).$$

This is a parabola in $\delta \geq 0$, with maximum occurring to the right of zero or at zero depending upon the sign of the R-average (the S-average will be positive with probability tending to one in the parameter region of interest). Accordingly

$$\sqrt{n}\widehat{\gamma} = \widehat{\delta} = \frac{1}{\sqrt{n}}\sum_{i=1}^n R\big((y_i-\widehat{\xi})/\widehat{\sigma}\big)\Big/\frac{1}{n}\sum_{i=1}^n S\big((y_i-\widehat{\xi})/\widehat{\sigma}\big) + O_p(n^{-1/2})$$

provided nominator is positive, and $\widehat{\delta} = 0$ if nominator is negative. But

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} R\big((y_i-\widehat{\xi})/\widehat{\sigma}\big) &= \frac{1}{n}\sum_{i=1}^{n} R\big(z_i-(\widehat{\xi}-\xi_0)/\sigma_0-\{(y_i-\xi_0)/\sigma_0^2\}(\widehat{\sigma}-\sigma_0)\big)+O_p(n^{-1})\\ &= \frac{1}{n}\sum_{i=1}^{n} R(z_i)-\frac{1}{n}\sum_{i=1}^{n} R'(z_i)(\widehat{\xi}-\xi_0)/\sigma_0\\ &\qquad -\frac{1}{n}\sum_{i=1}^{n} R'(z_i)z_i(\widehat{\sigma}-\sigma_0)/\sigma_0+O_p(n^{-1}),\end{aligned}$$

where $z_i = (y_i-\xi_0)/\sigma_0$, and similarly for the S-function term. Judicious calculations based on this show that $\frac{1}{n}\sum_{i=1}^{n} S\big((y_i-\widehat{\xi})/\widehat{\sigma}\big)$ goes to $7/2$ and that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} R\big((y_i-\widehat{\xi})/\widehat{\sigma}\big) &= \sqrt{n}\,\bar{W}_n-(2/\sigma_0)\sqrt{n}(\widehat{\sigma}-\sigma_0)+O_p(n^{-1/2})\\ &= \tfrac{7}{3}\sqrt{n}\,\bar{W}_n-\tfrac{7}{3}\sigma_0\sqrt{n}\,\bar{V}_n+O_p(n^{-1/2})\end{aligned}$$

on the set Ω_n. This finally means that $\sqrt{n}\widehat{\gamma} = \Delta_n + O_p(n^{-1/2})$ in the $\Delta_n + O_p(n^{-1/2}) > 0$ case and is 0 in the $\Delta_n + O_p(n^{-1/2}) \le 0$ case. This proves what was needed. □

Reference

Hjort, N.L. (1991). Estimation in moderately misspecified models. Technical report, University of Oslo; submitted for publication.